Assignment 2 Advance Programming
Customer Segmentation Using Clustering

Introduction

In this project, we conducted customer segmentation for an e-commerce dataset. The data included customer, product, order, payment, seller, and product category information. We merged these datasets into a comprehensive dataframe for analysis.

Selected Attributes

For further segmentation, we focused on the following attributes:
- order_status
- customer_state
- order_item_id
- price
- freight_value
- payment_sequential
- payment_type
- payment_installments
- payment_value
- order_purchase_year
- order_purchase_month
- order_purchase_day
- order_purchase_hour

Data Preprocessing

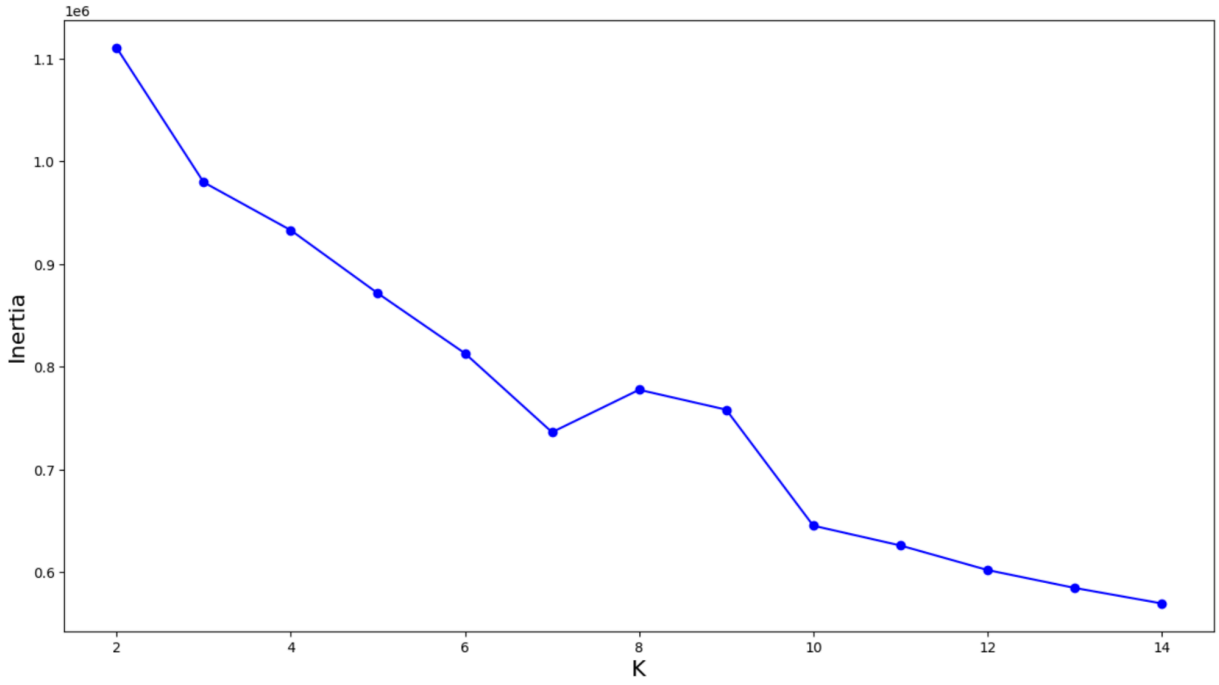
We applied preprocessing techniques to prepare the data for analysis. Categorical attributes were converted into one-hot encoded format, while numerical variables underwent z-score normalization to standardize their scales.

Dimensionality Reduction with PCA

We performed Principal Component Analysis (PCA) to reduce the dimensionality of the dataset while preserving 95% of the variance. This step was essential for minimizing the complexity of the dataset while retaining important information.

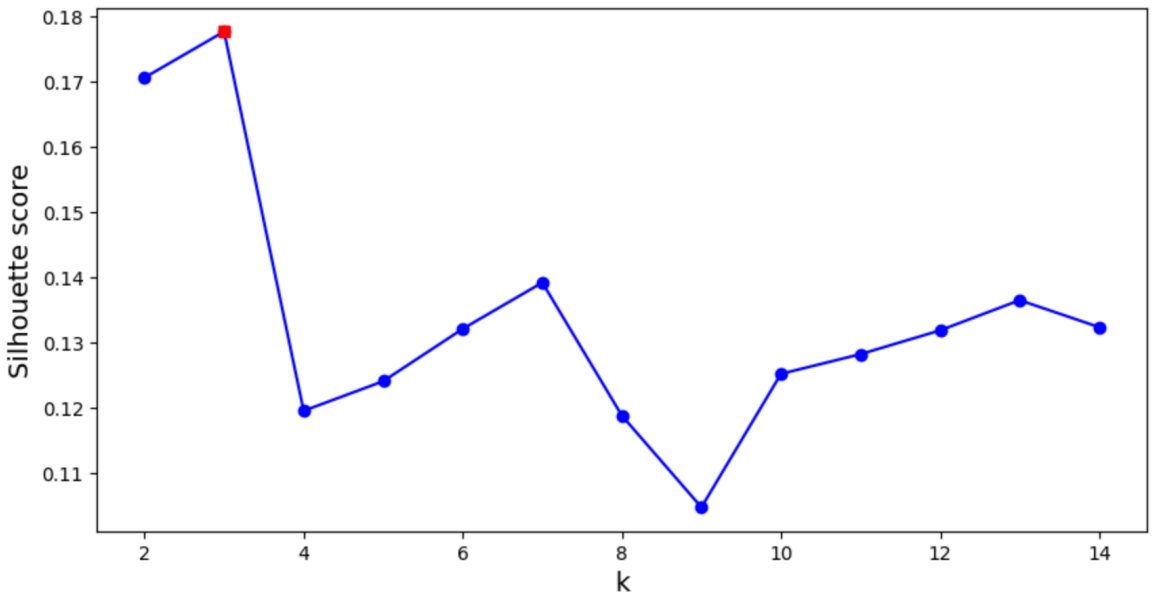
K-Means Clustering

K-Means clustering was applied to the reduced dataset, with k values ranging from 2 to 15. We calculated the inertia for each k value to assess the compactness of the clusters. The inertia graph is included.



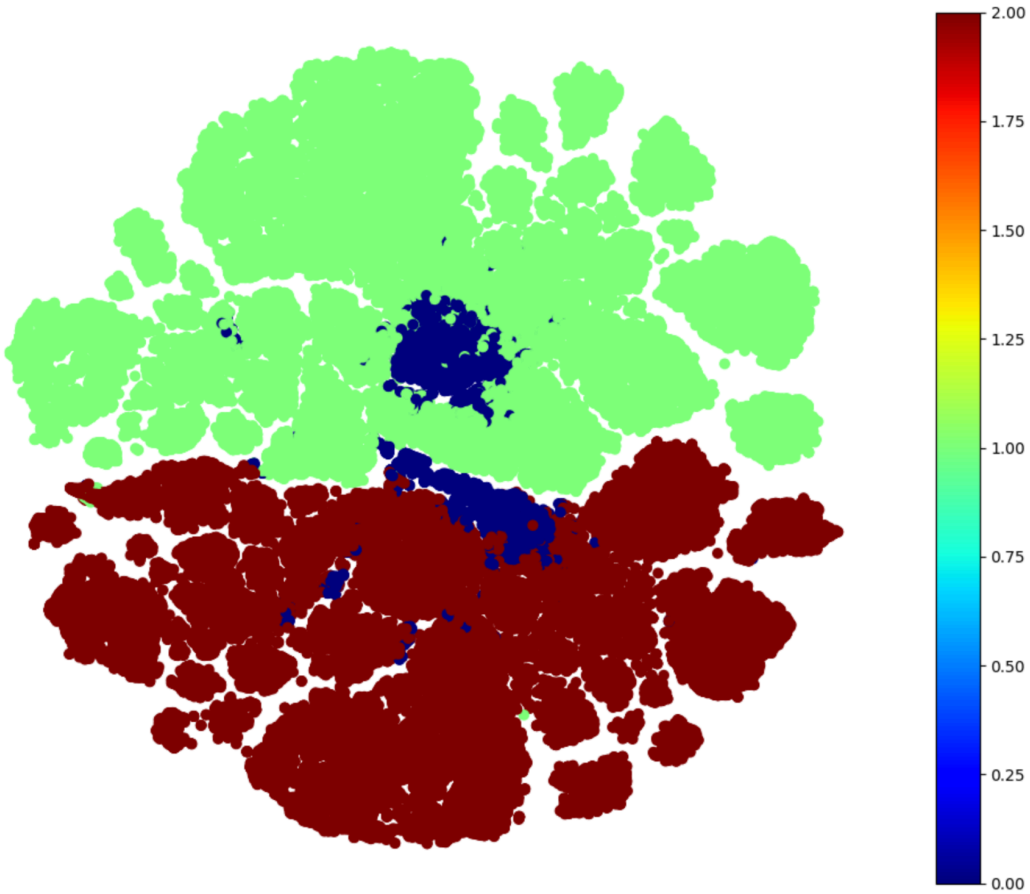
Silhouette Score Calculation

We computed silhouette scores for each k value to evaluate the quality of the clustering. This metric measures how well-separated the clusters are, providing insight into the optimal number of clusters. Based on the silhouette scores, the best value for k was determined to be 3. The silhouette score graph is as follows



Visualization with t-SNE

To visualize the clusters, we applied t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the data to two dimensions. This allowed us to effectively illustrate the distinct customer segments identified by the clustering process.


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00